Lending Club Case Study

Exploratory Data Analysis

Abhishek Dixit

Adesh Pathak

- Problem Statement
- Data Summary
- Data Cleaning
- Data conversions And Derived Columns
- Dropping/Imputing the Rows
- Univariate Analysis
- Bivariate Analysis
- Conclusions

Problem Statement

Problem:

- You work for a consumer finance company which specialises in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
    - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
    - If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Objective:

- Use EDA to understand how consumer attributes and loan attributes influence the tendency of default

Constraints:

- When a person applies for a loan, there are two types of decisions that could be taken by the company:
    - **Loan accepted**: If the company approves the loan, there are 3 possible scenarios described below:
        - **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
        - **Current:** Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
        - **Charged-off**: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan
    - **Loan rejected:** The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the
        - loan was rejected, there is no transactional history of those applicants with the company and so this data is notavailable with the company (and thus in this dataset)

3

Data Summary

- Loan.csv file contains 39717 rows and 111 columns.
- There is no sub-headers or sub-Footer in the given data set.

Data Cleaning

- There were no duplicates rows found.
- There were 1140 rows present of loan_status, ‘current’ which has been deleted as loan_status ‘current’ does n’t participate in analysis.
- There were 55 columns which is having all the rows values as null/blank and doesn’t participate in analyse has been removed.
- ‘mths_since_last_record’ and ‘mths_since_last_delinq’ has more than 60% has null values so these columns has been deleted.
- ‘desc’ and ‘title’ text/description values and doesn’t participate has been dropped from analysis.
- 11 columns whose values were 1, and is uniqueness in nature has been dropped from analysis.
- ‘id’, ‘member_id’, ‘url’, ‘zip_code’, are unique in nature and ‘funded_amnt_inv’, ‘total_pymnt_inv’ are same as ‘funded_amnt’, ‘total_pymnt’ so these columns has been deleted.
- After all the Data cleaning process we are left with 35367 rows and 36 columns.

5

Data Conversions and Derived Columns

- Additional string value has been trimmed from ‘term’ column and has been converted to int data types.
- ‘int_rate’ and ‘revol_util’ has been converted from string to int. Additional ‘%’ has been trimmed.
- Column ‘loan_funded_amnt’ and ‘funded_amnt’ converted to float.
- ‘emp_length’ has been converted to number from 0(represent 0-1 year) to 10(represent greater than 9 years).
- ‘issue_d’, ‘last_pymnt_d’, ‘last_credit_pull_d’ , ‘earliest_cr_line’ have been converted to date type.
- Creating derived columns for ‘issue_year’ and ‘issue_month ’ from ‘issue_d’, ‘last_pymnt_d_month’ and ‘last_pymnt_d_year’ from ‘last_pymnt_d’, ‘last_credit_pull_d_month’ and ‘last_credit_pull_d_year’ from ‘last_credit_pull_d’, ‘earliest_cr_line_month’ and ‘earliest_cr_line_year’ from ‘earliest_cr_line’.
- ‘profit_loss’ column is created for each borrower which represent whether the company gain profit (‘+’ sign) or lose (‘-’ sign) their money.
- ‘ratio’ column is derived which represent the ratio of borrower annual income to the loan amount.

Dropping/Imputing the rows

- ‘emp_title’,‘emp_lenght’ and pub_rec_bankruptcies contains 6.18%, 2.67% and 1.80% of rows as null, which is very small percentage of data which we candrop it.
- Total % of rows deleted: 8.32%,
- Outliers exits for numeric data 'loan_amnt', 'funded_amnt','int_rate', 'installment', ‘annual_inc' so these are removed while analyzing.
- Outliers’ treatment has been done for above fields using Upper Fence and Lower Fence mechanism.

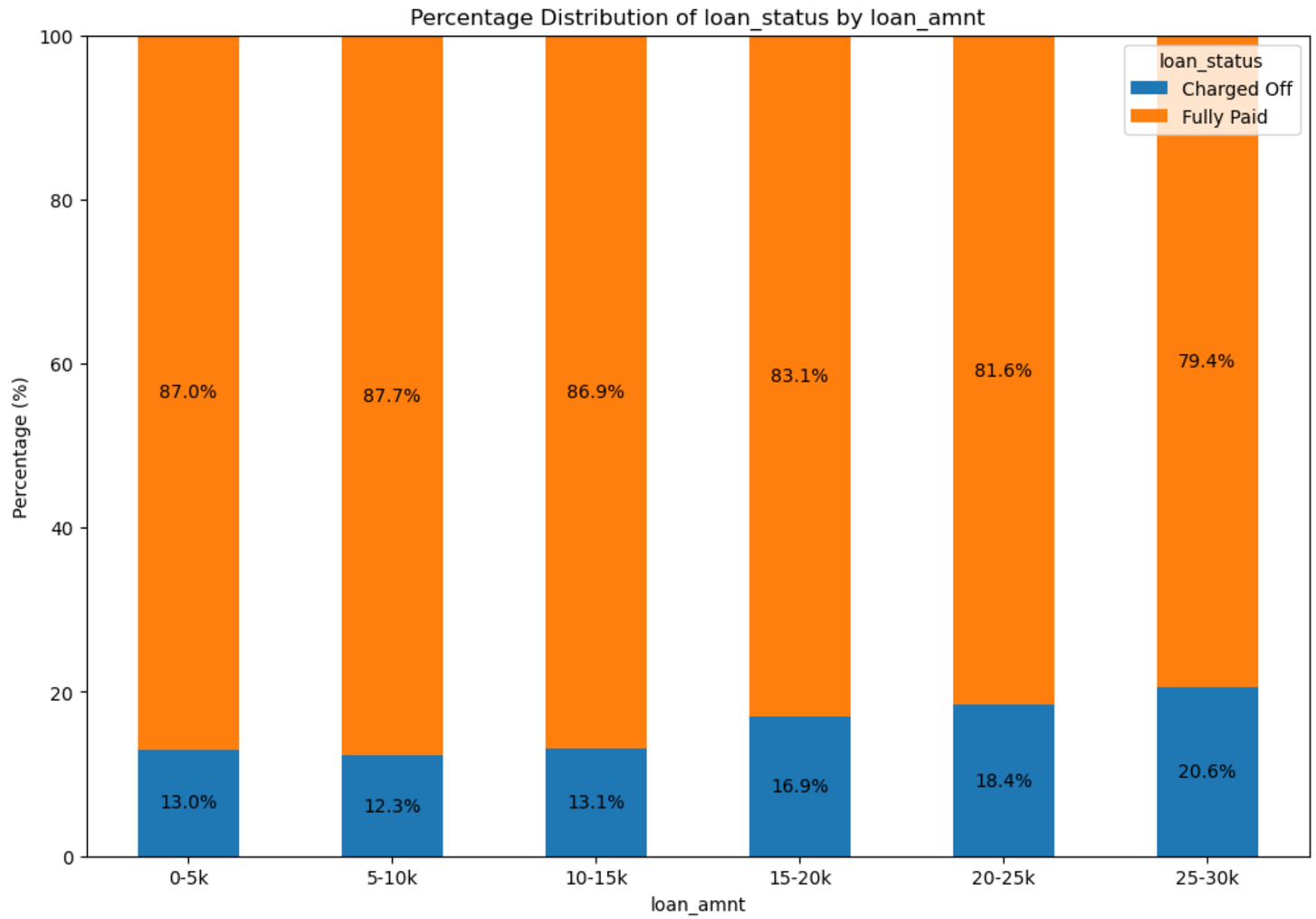
Univariate And Segmented Univariate Analysis

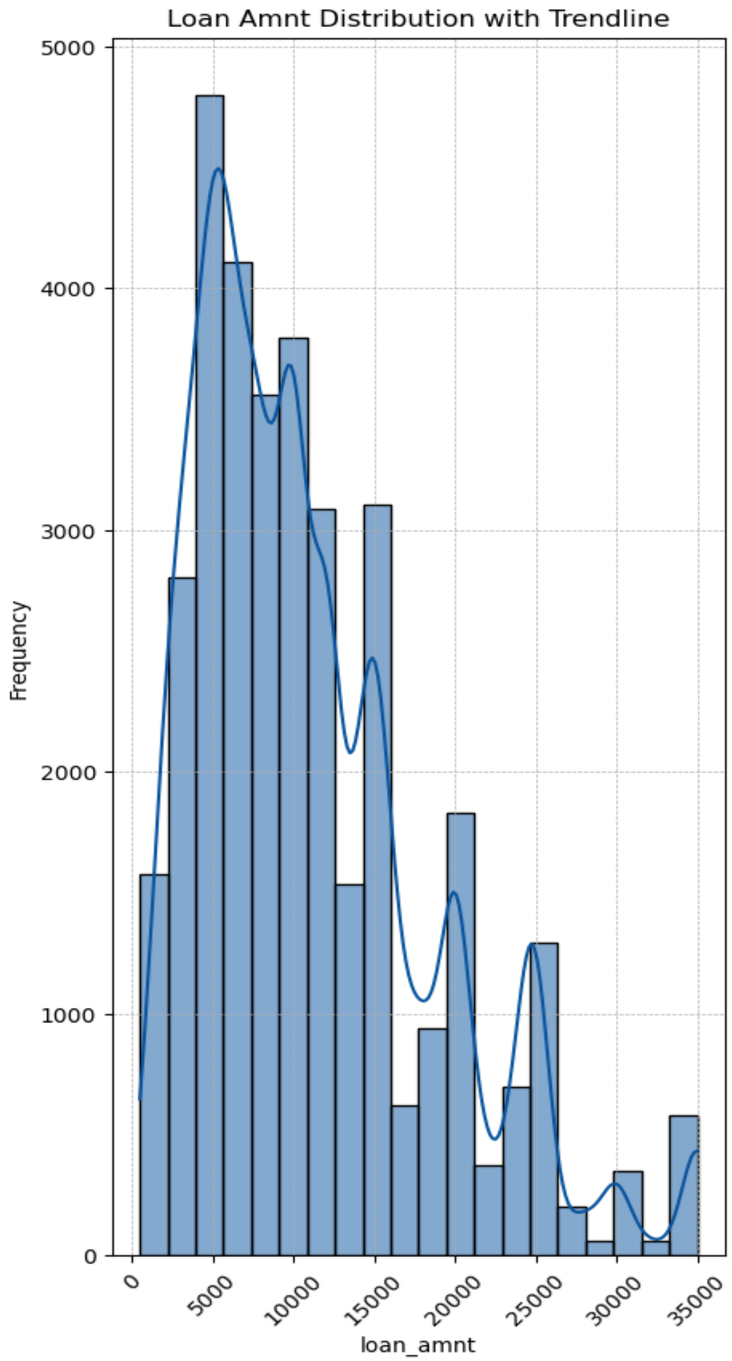
7

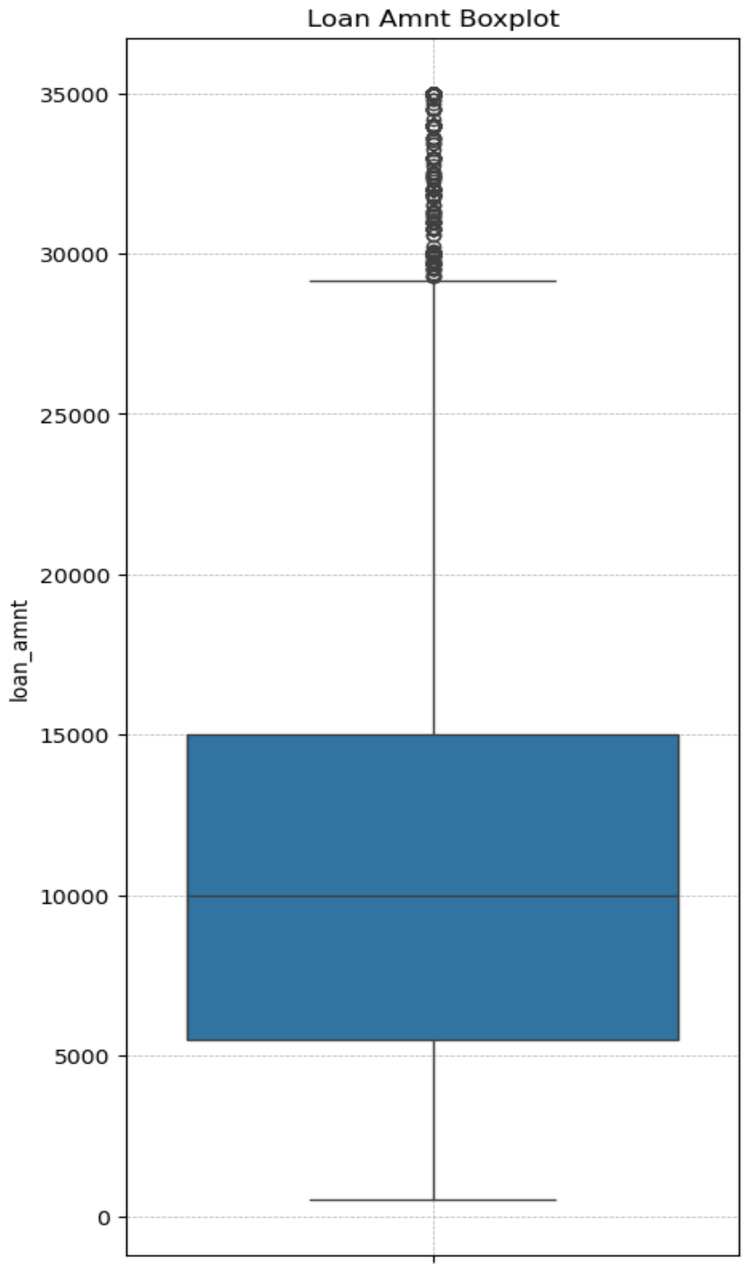
Loan Status & Loan Amount

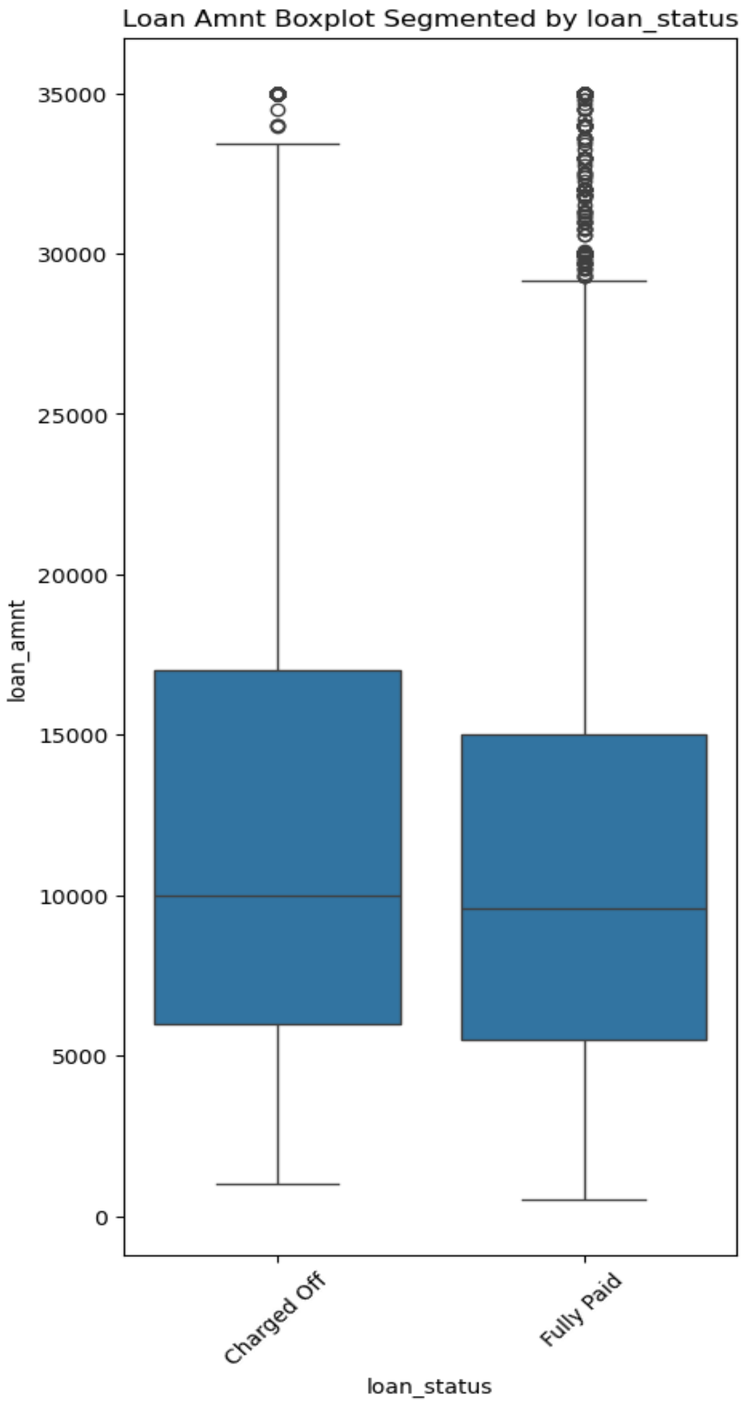
Observations:

- Approximate 14% borrower are defaulted in the data set.
- Overall loan amount varies from 500 to 35000.
- Charged off loans have high std, mean, median than fully paid.
- There are Outliers in the Loan Amount So we remove the outliers First
- After removing outliers, we have bucketed the loan amount in the gap of 5000
- Most of the loan amount are between 5000 to 10000
- Fully Paid loans have low IQR in comparison to charged off.
- Charged Off has high average and std in comparison to Fully Paid
- % default increasing with increase in loan amount range

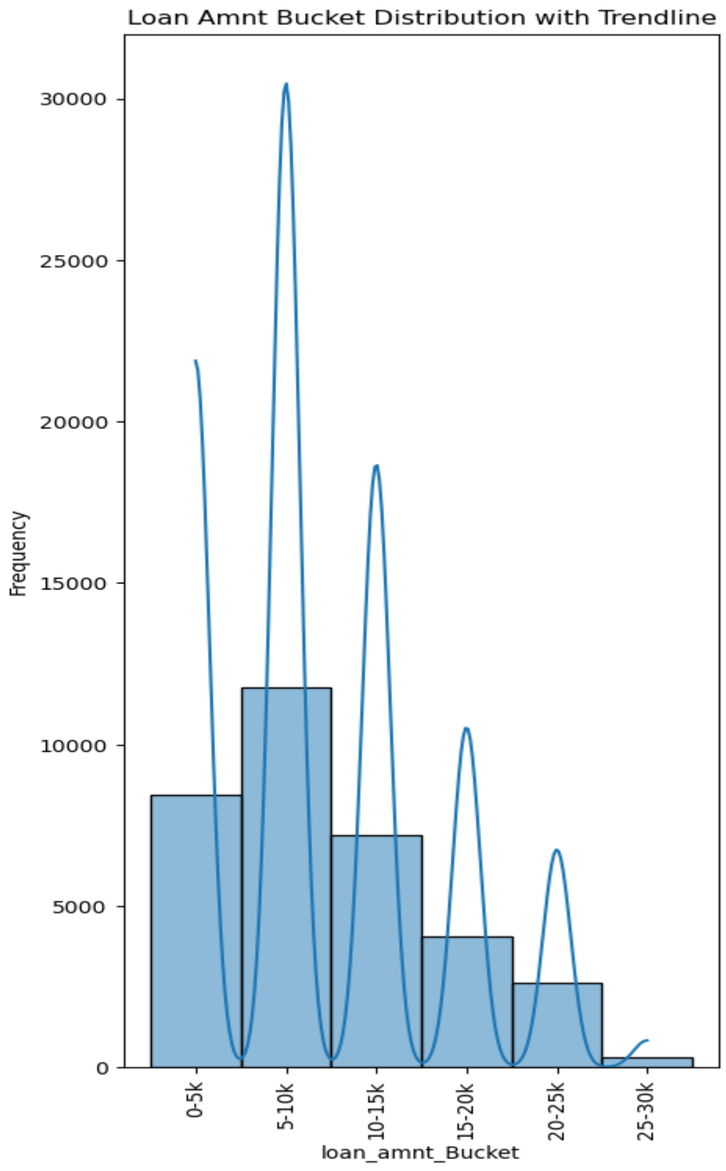


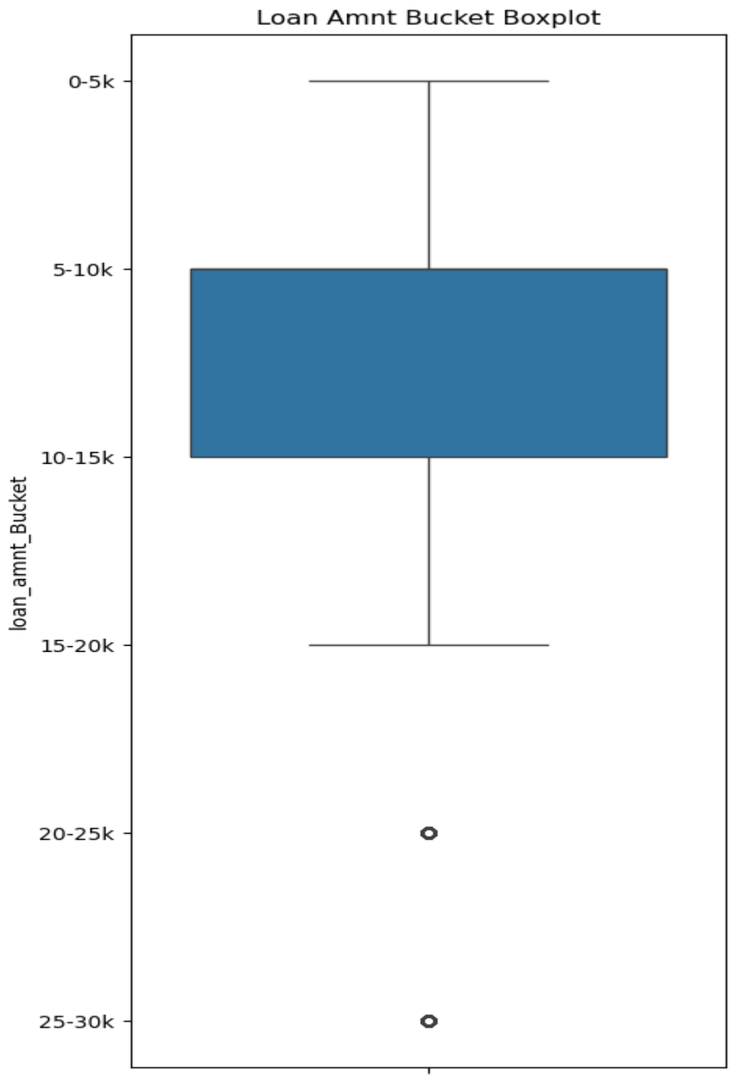

Loan Amnt Distribution with Trendline
Frequency
loan_amnt
0
1000
2000
3000
4000
5000
0
5000
10000
15000
20000
25000
30000
35000

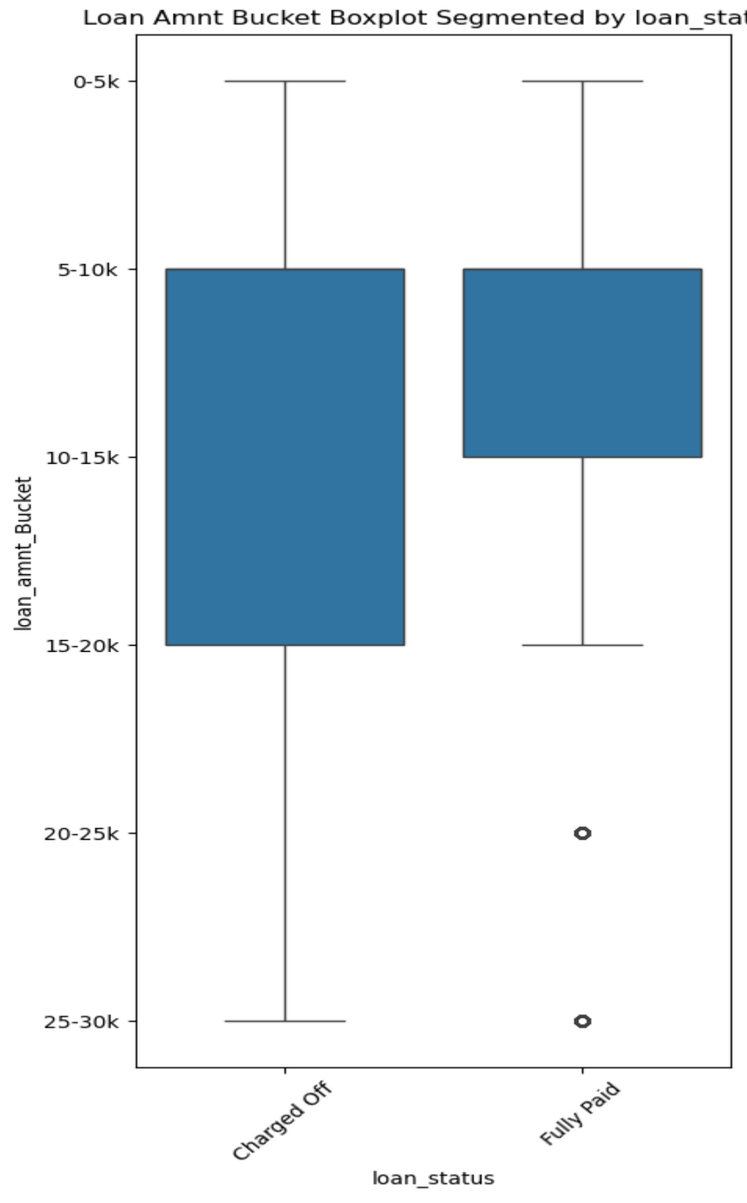

Loan Amnt Boxplot
loan_amnt
0
5000
10000
15000
20000
25000
30000
35000


Loan Amnt Boxplot Segmented by loan_status
loan_amnt
loan_status
Charged Off
Fully Paid
0
5000
10000
15000
20000
25000
30000
35000

9


Loan Amnt Bucket Distribution with Trendline
30000
25000
20000
15000
10000
5000
0
Frequency
0-5k
5-10k
10-15k
15-20k
20-25k
25-30k
loan_amnt_Bucket

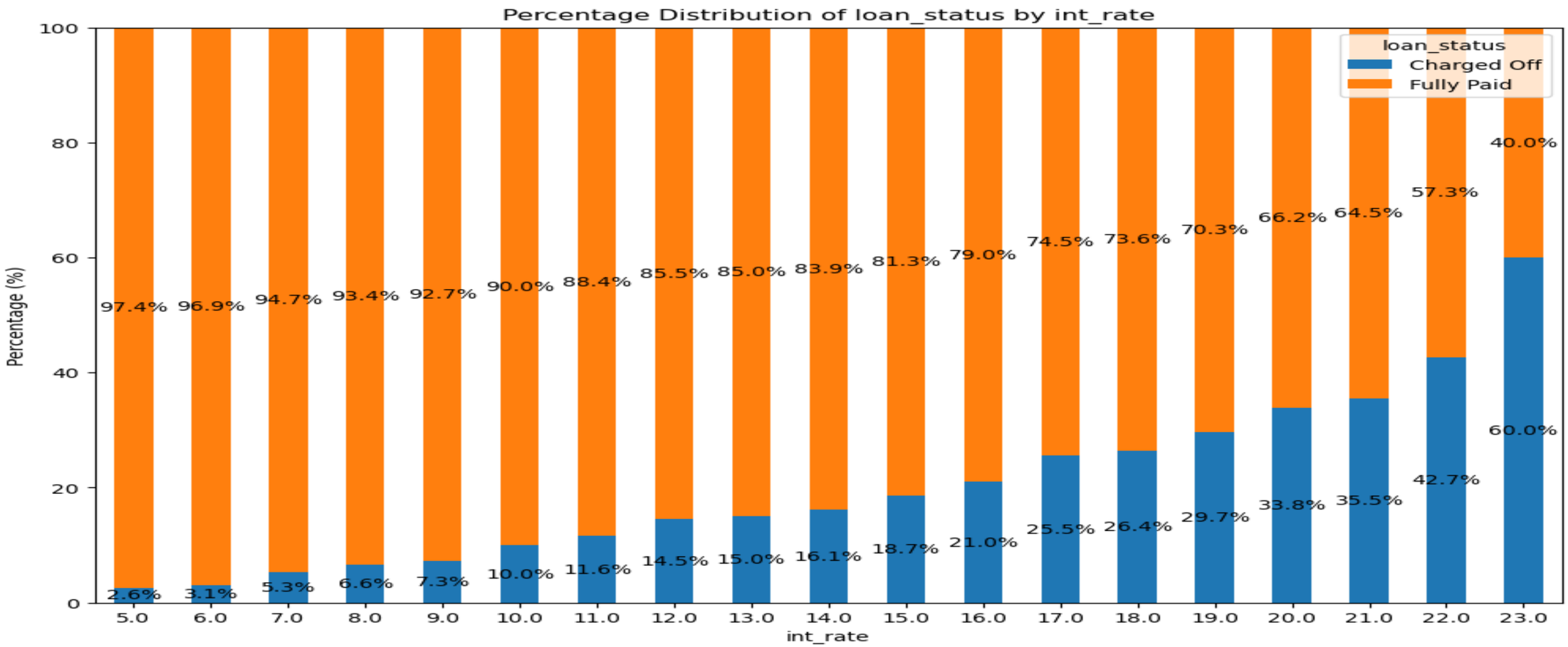

Loan Amnt Bucket Boxplot
0-5k
5-10k
10-15k
15-20k
20-25k
25-30k
loan_amnt_Bucket

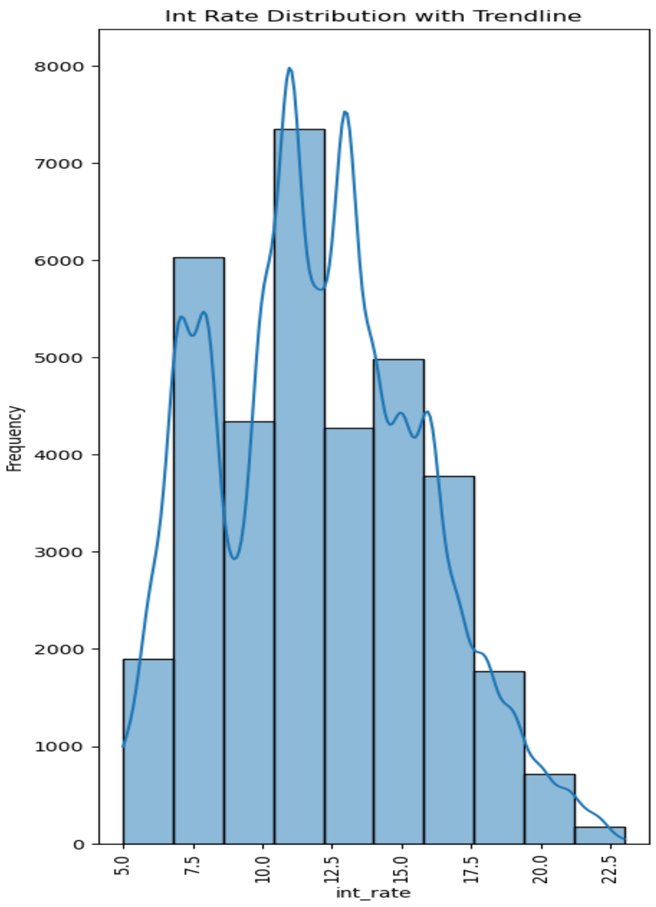

Loan Amnt Bucket Boxplot Segmented by loan_status
0-5k
5-10k
10-15k
15-20k
20-25k
25-30k
loan_amnt_Bucket
Charged Off
Fully Paid
loan_status

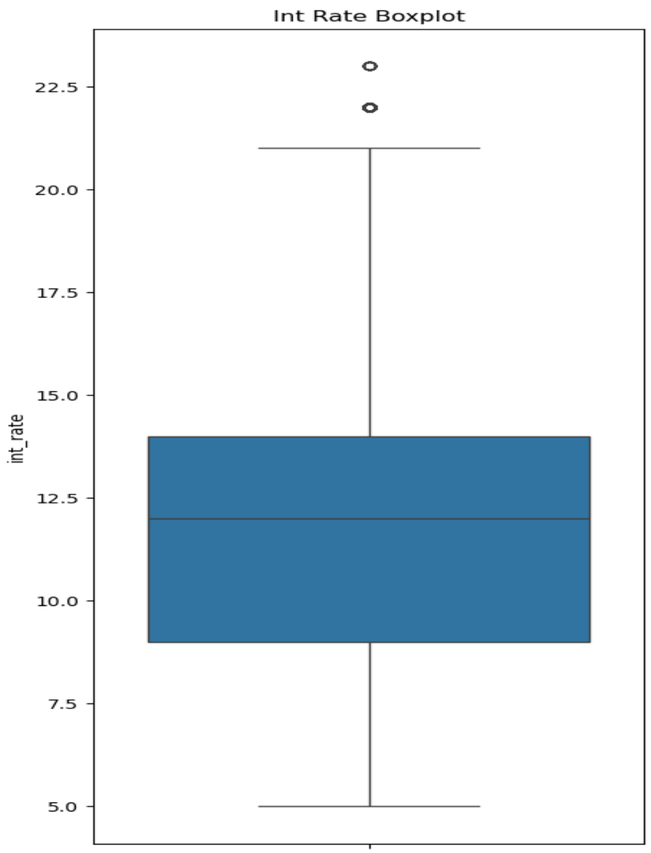
Interest Rate

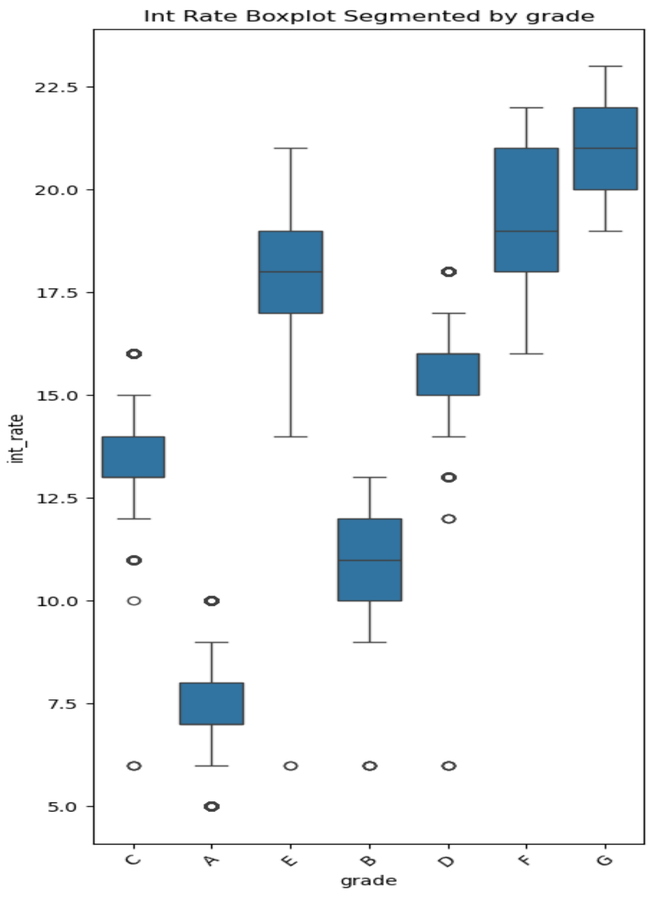
Observations:

- First of all, we have removed outliers.
- After removing outliers, we have round off interest rate in difference of 1.
- overall interest rate varies from 5% to 23% after removing outliers
- 11 and 13% interest rate are more frequent in complete data set.
- The interest rate for Charged Off loans appear to be higher than for Fully paid. This is naturally expected. As, the risk increases the rate of interest imposed on the loan also increases
- LC provided grade E,F,G grade has the highest rate of interest.
- on increasing interest rate, the chance of being default also increases.




Int Rate Distribution with Trendline
Frequency
8000
7000
6000
5000
4000
3000
2000
1000
0
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
int_rate

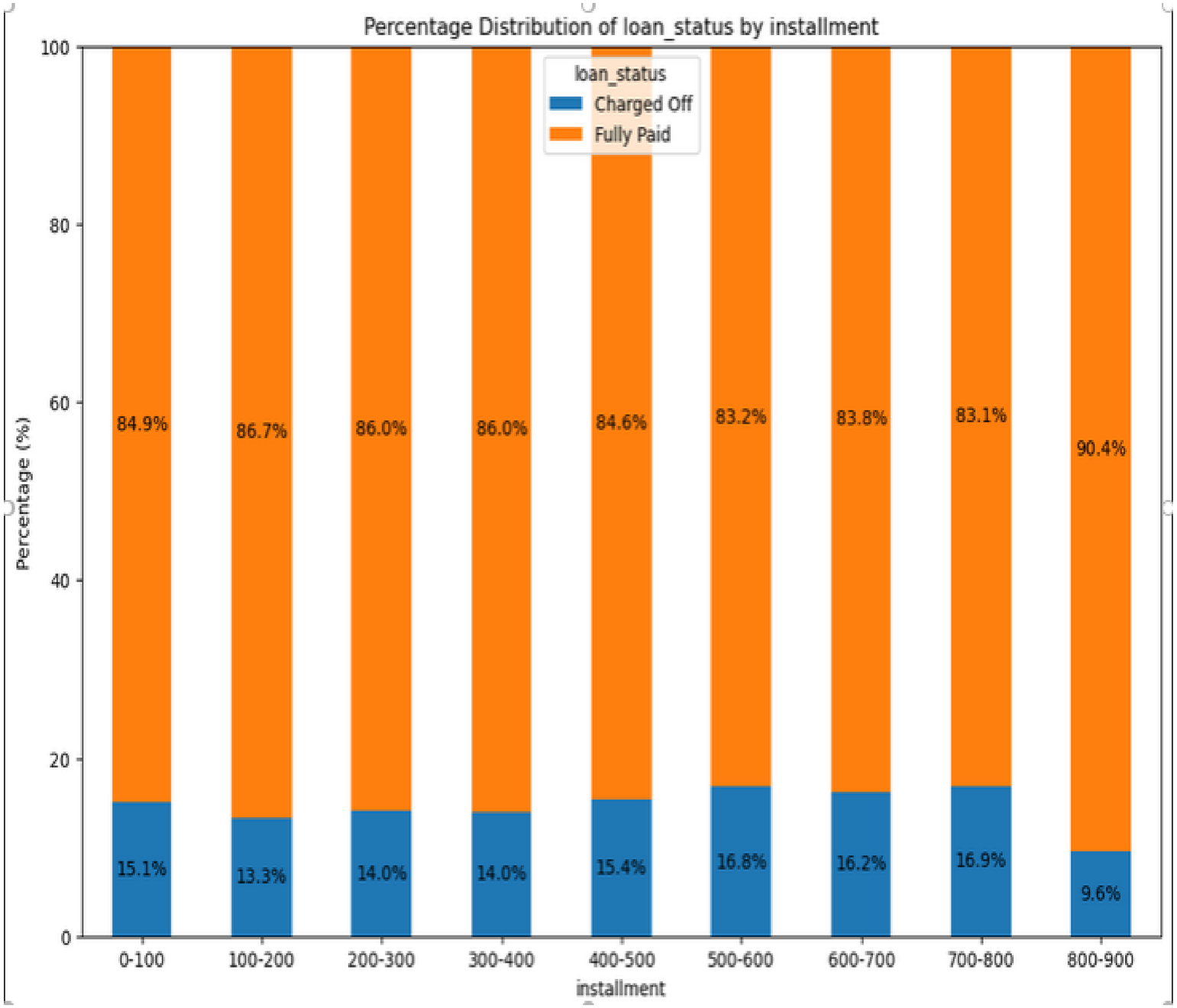

Int Rate Boxplot
int_rate
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0

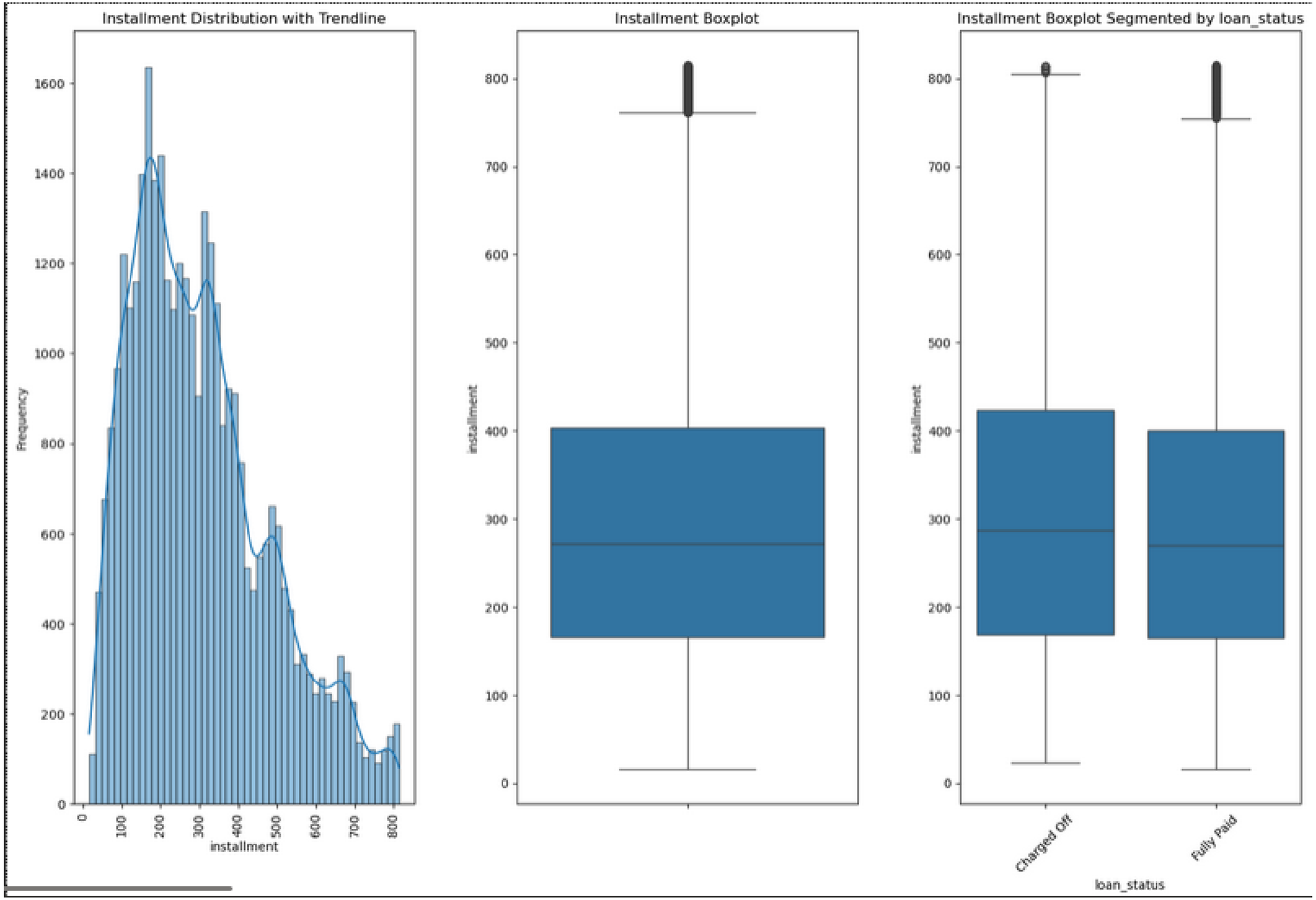

Int Rate Boxplot Segmented by grade
int_rate
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
C
A
E
B
D
F
G
grade

Installment

Observations:

- First of all, we have removed outliers.
- overall installment varies from 16.08 to 1305.19.
- Most of the Installment fall between 0 to 400.
- loans Charged Off have high installment on average

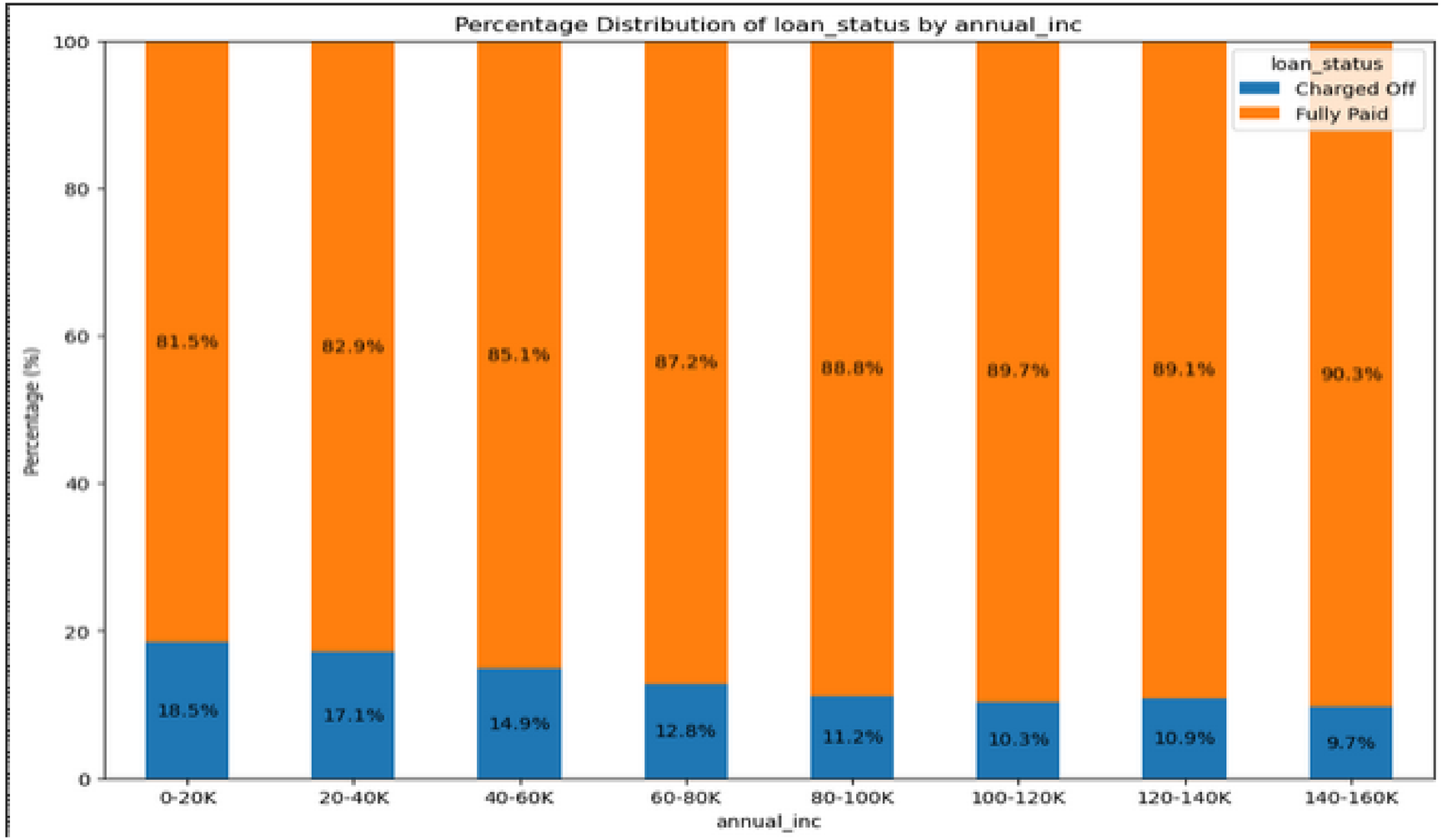



Installment Distribution with Trendline
Frequency
installment
0
200
400
600
800
1000
1200
1400
1600
0
100
200
300
400
500
600
700
800
Installment Boxplot
installment
0
100
200
300
400
500
600
700
800
Installment Boxplot Segmented by loan_status
installment
0
100
200
300
400
500
600
700
800
Charged Off
Fully Paid
loan_status

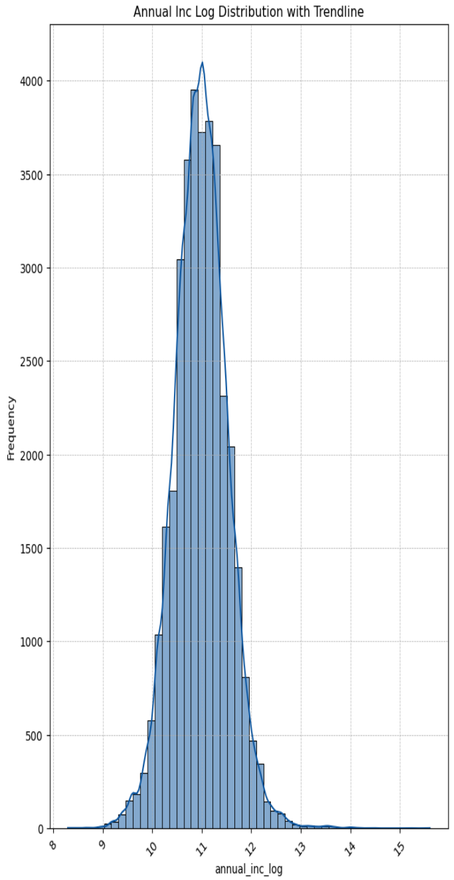
Annual Income

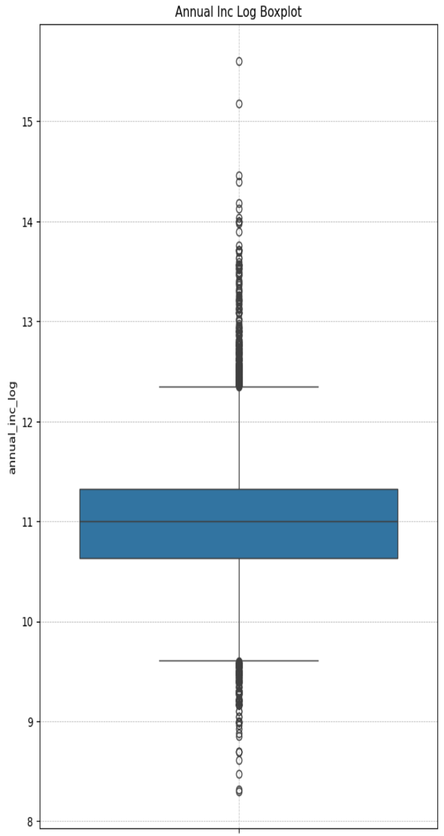
Observations:

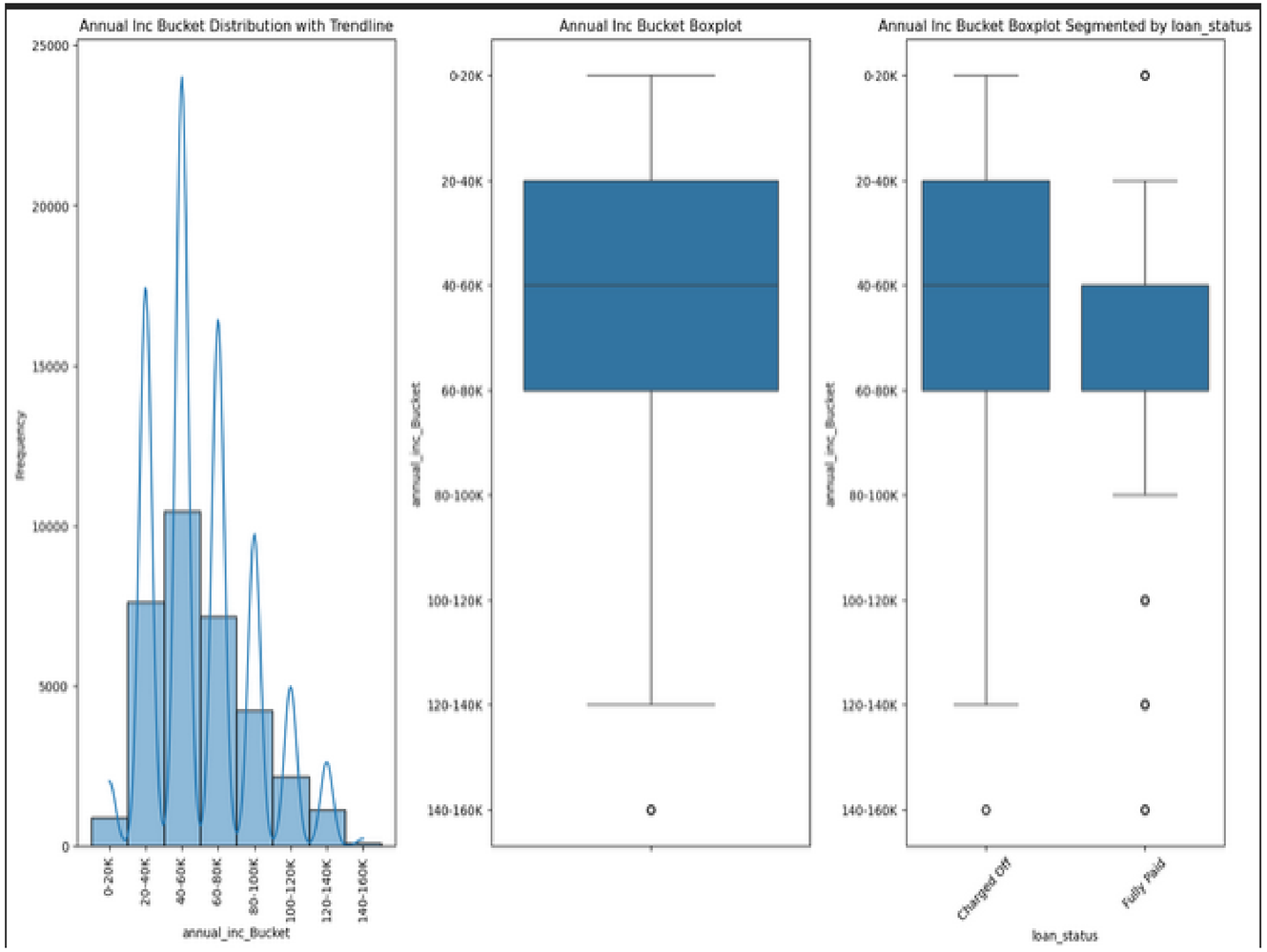
- The Log plot of annual income show normal distribution means symmetric about it means.
- To further Analysis we have remove outliers.
- After removing outliers, we have created bucket of 20K difference.
- most of the people in the data set fall in income range 30-70k after removing outliers
- Loan defaults are higher for lower income, and progressively reduce as incomes go up rate of interest imposed on the loan also increases
- On lower Salery bucket range higher the chance of default.



15


Annual Inc Log Distribution with Trendline
Frequency
annual_inc_log

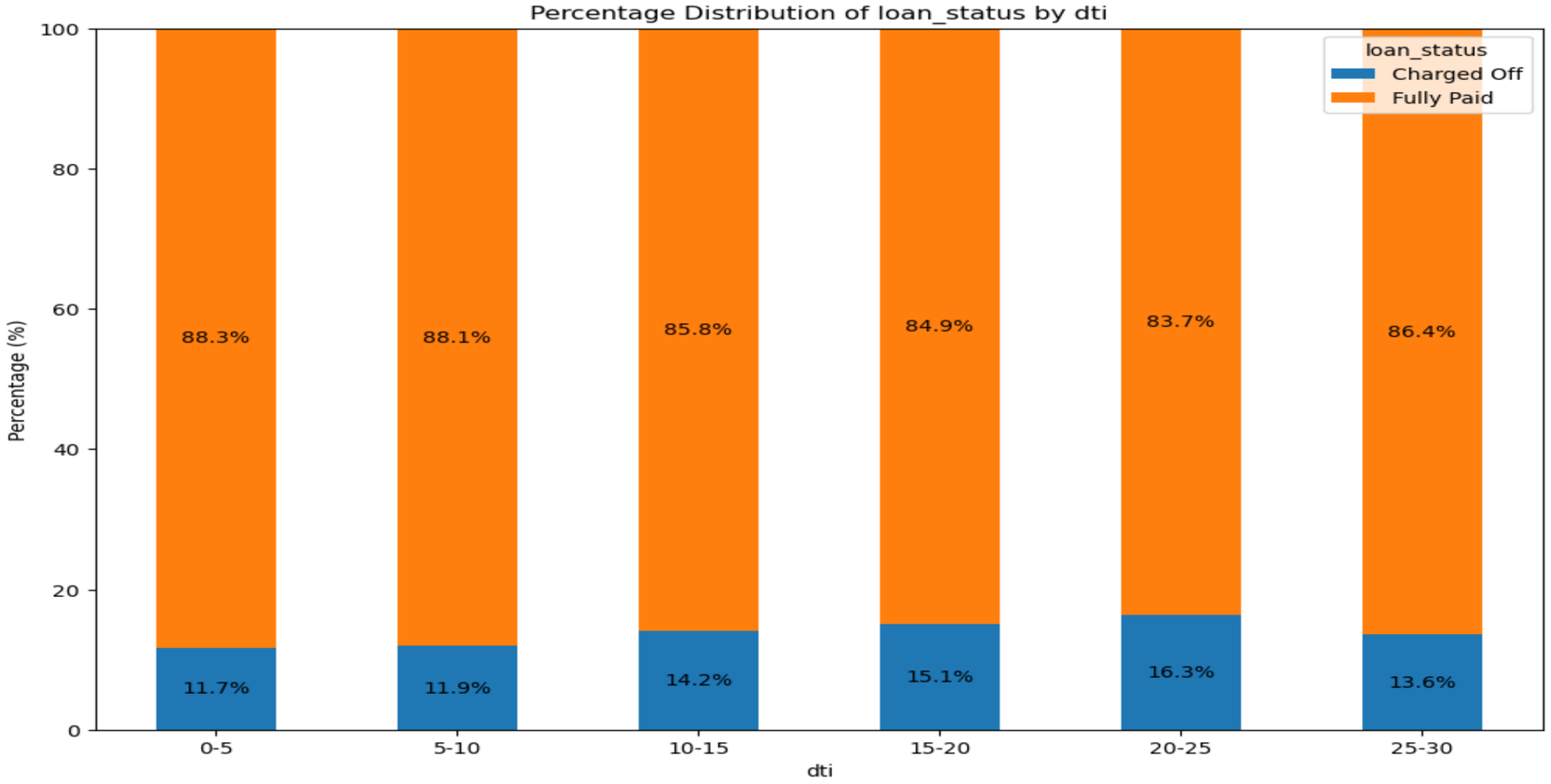

Annual Inc Log Boxplot
annual_inc_log


Annual Inc Bucket Distribution with Trendline
Frequency
annual_inc_Bucket
Annual Inc Bucket Boxplot
annual_inc_Bucket
Annual Inc Bucket Boxplot Segmented by loan_status
annual_inc_Bucket
Charged Off
Fully Paid
loan_status

DTI(Debt to Income Ratio)

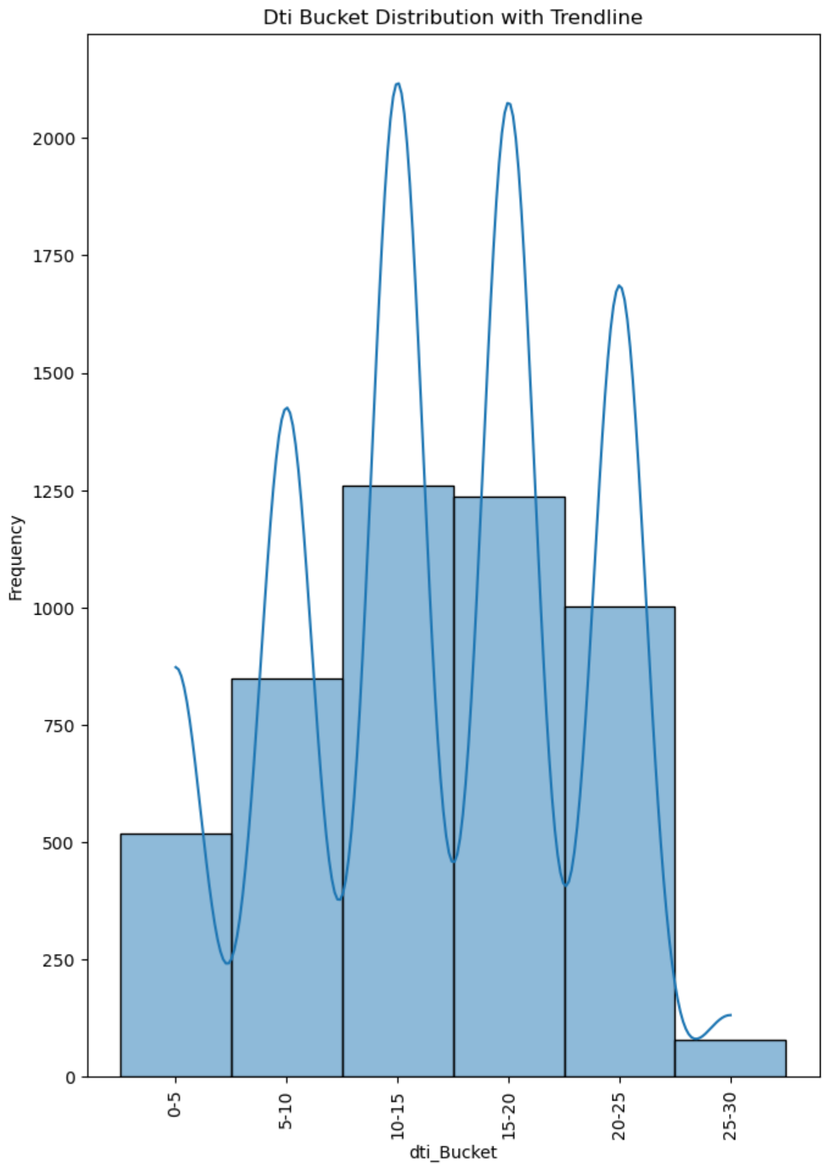
Observations:

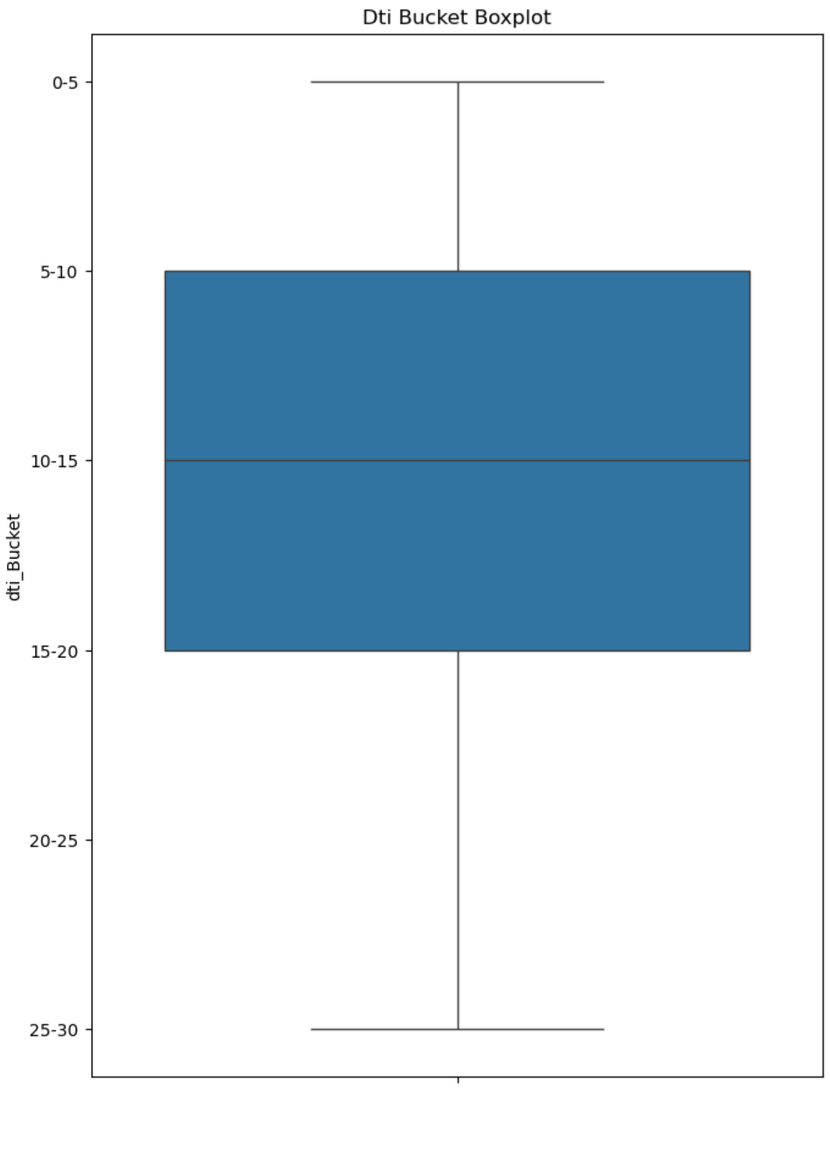
- First of all, we have removed outliers.
- After removing outliers, we have created bucket of 5-5 difference.
- The dti index varies from min 0 max of 30. The median dti is of 13.5
- Dti value has high frequency between 0-25.
- As the debt-to-income ratio increases the percent of being default of the borrower also increase.



to

17

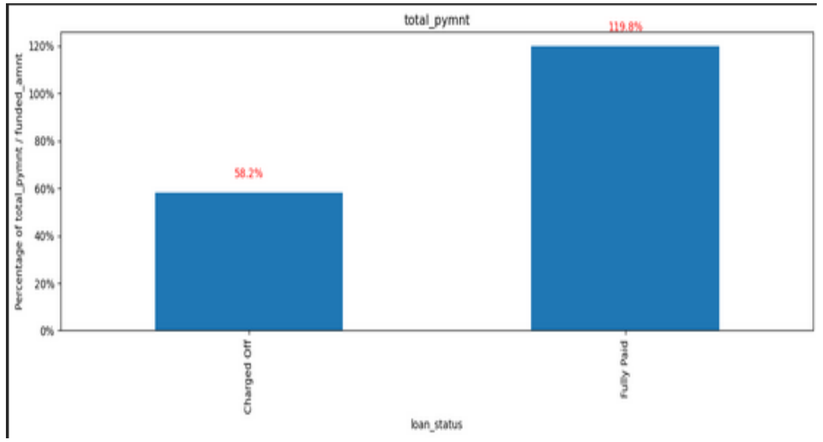

Dti Bucket Distribution with Trendline
Frequency
0
250
500
750
1000
1250
1500
1750
2000
0-5
5-10
10-15
15-20
20-25
25-30
dti_Bucket

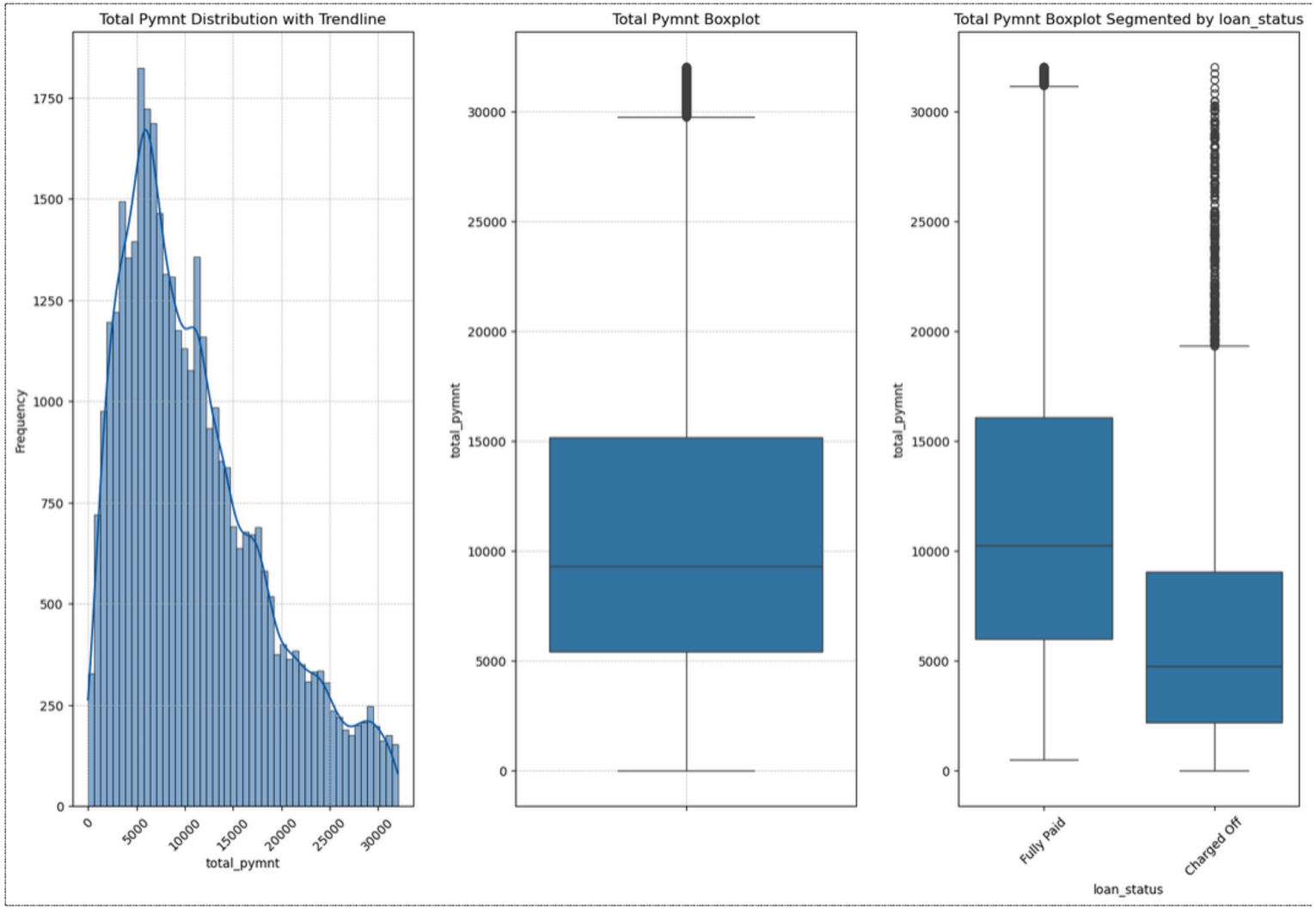

Dti Bucket Boxplot
dti_Bucket
0-5
5-10
10-15
15-20
20-25
25-30

Total Payment

Observations:

- First of all, we have removed outliers.
- The average payment received to date for the Charged Off loan is comparatively less than Fully Paid loans
- Without removing outliers in total payment LC get approximate 59% of its payment in charged off and in fully paid it gets 20% profit.



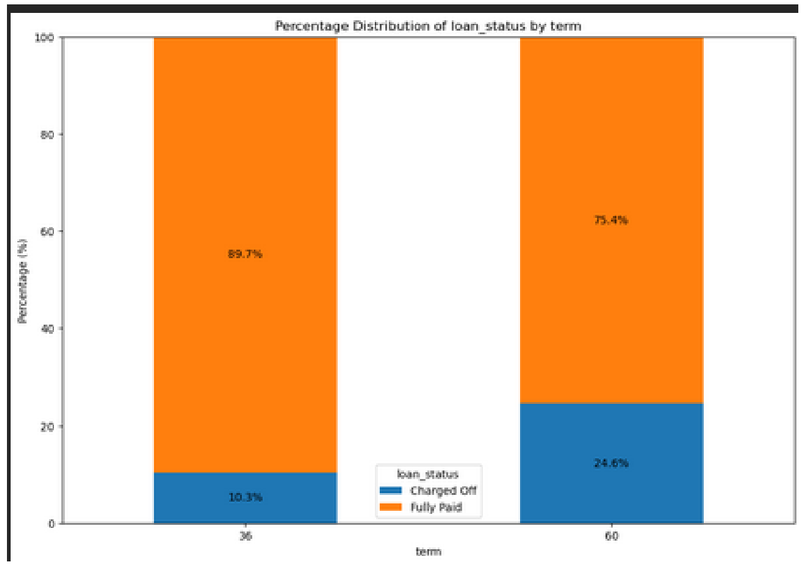


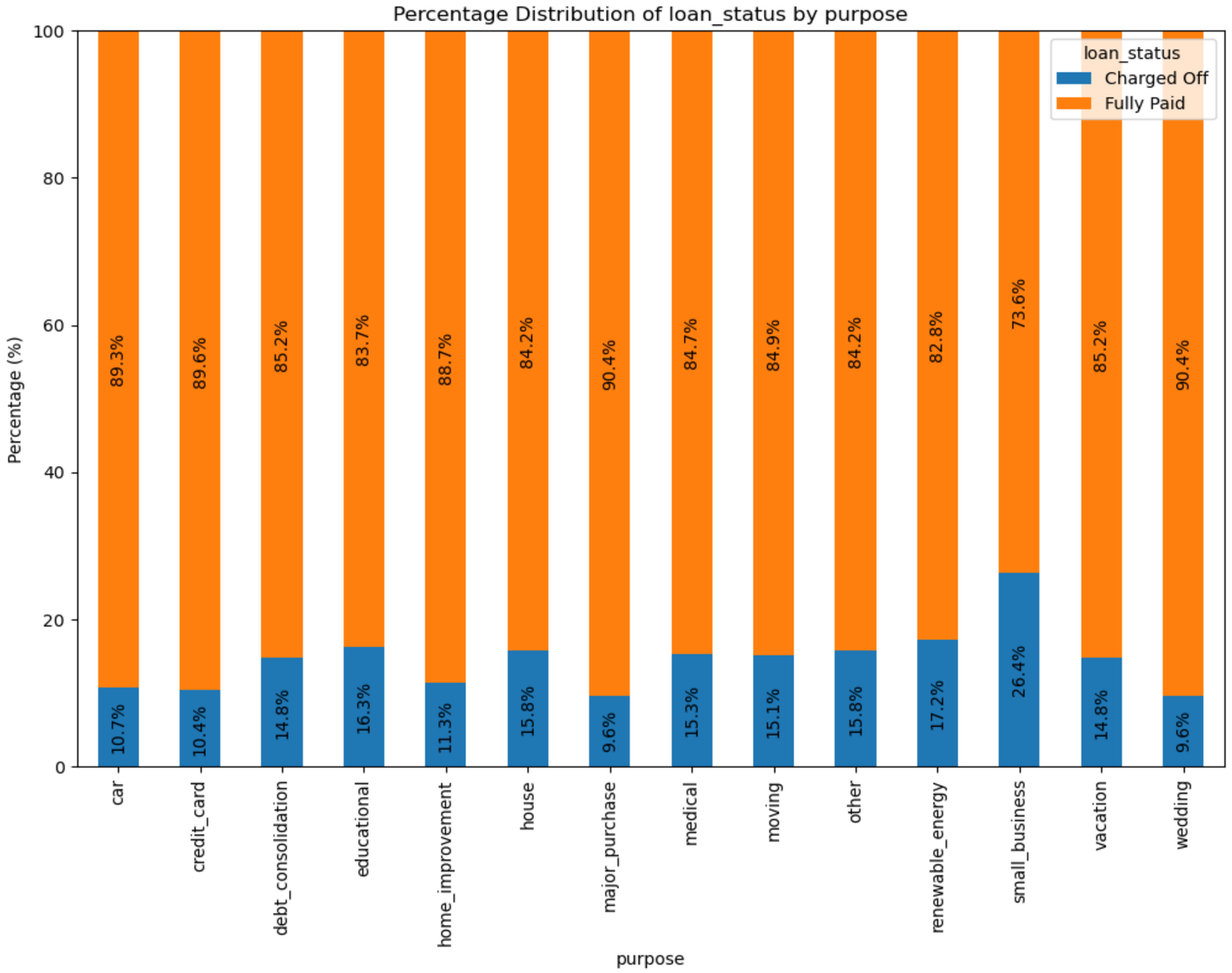
19

Term and Purpose

Observations:

- Higher the term higher chance of default.
- Small business has higher chance of default.

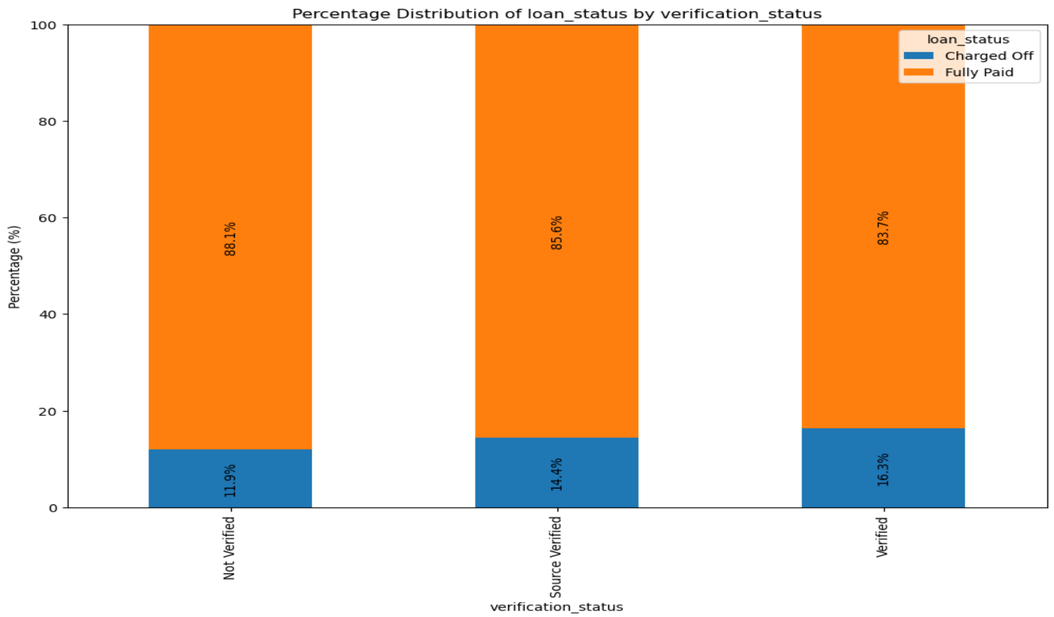


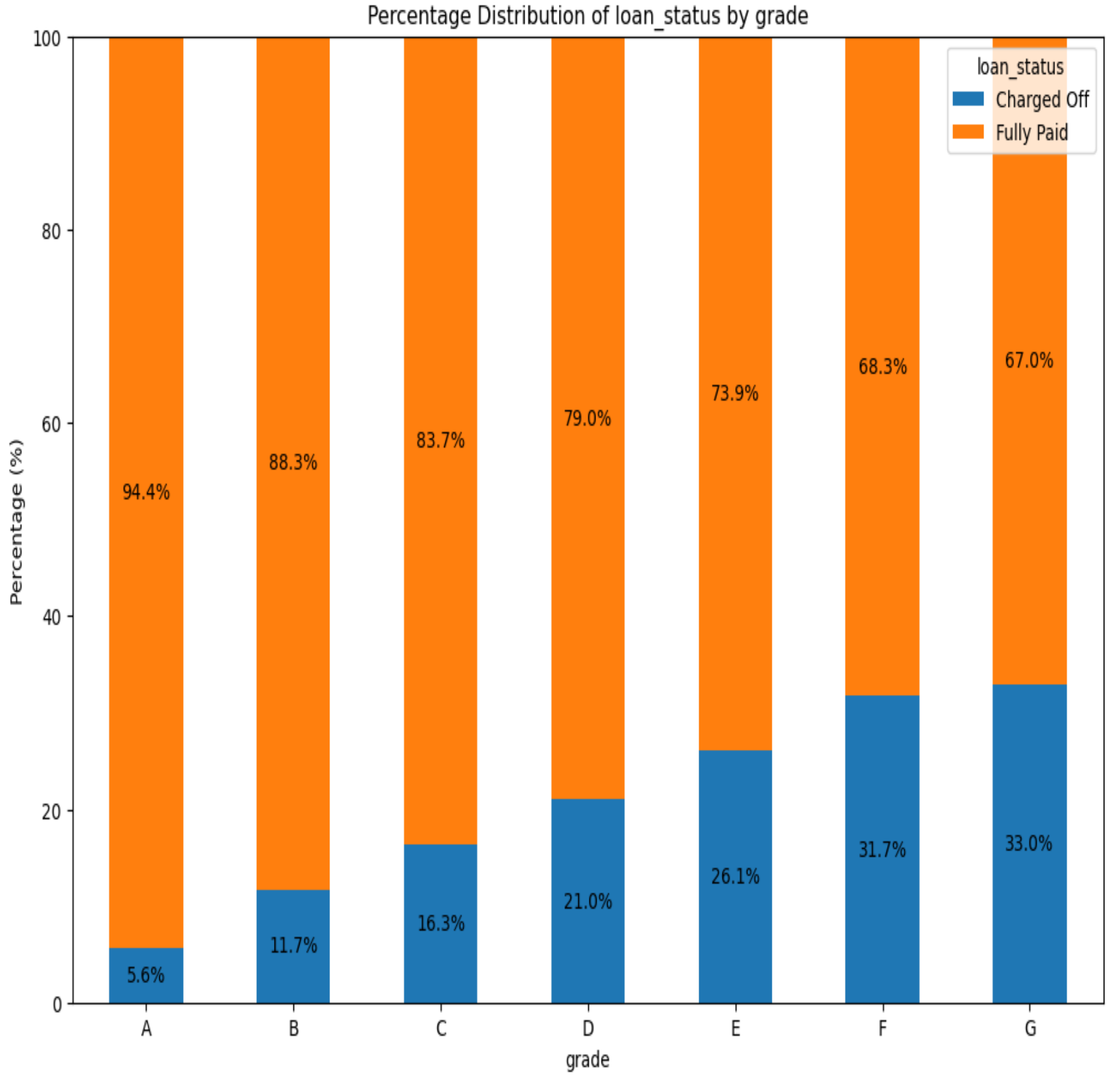


Grade and Verification Status

Observations:

- As the grade provided by lending club increases the chance of being default also increases.
- Source verified and verified has higher % of default



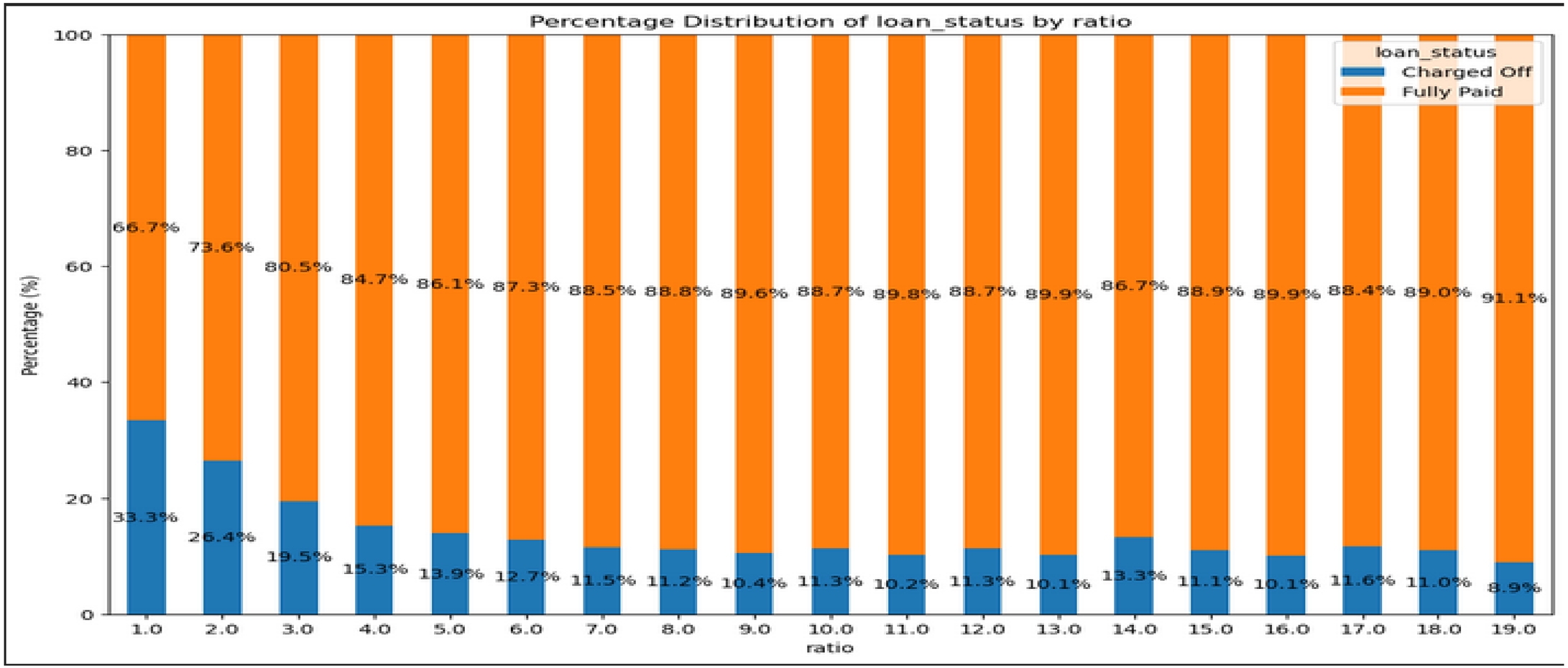


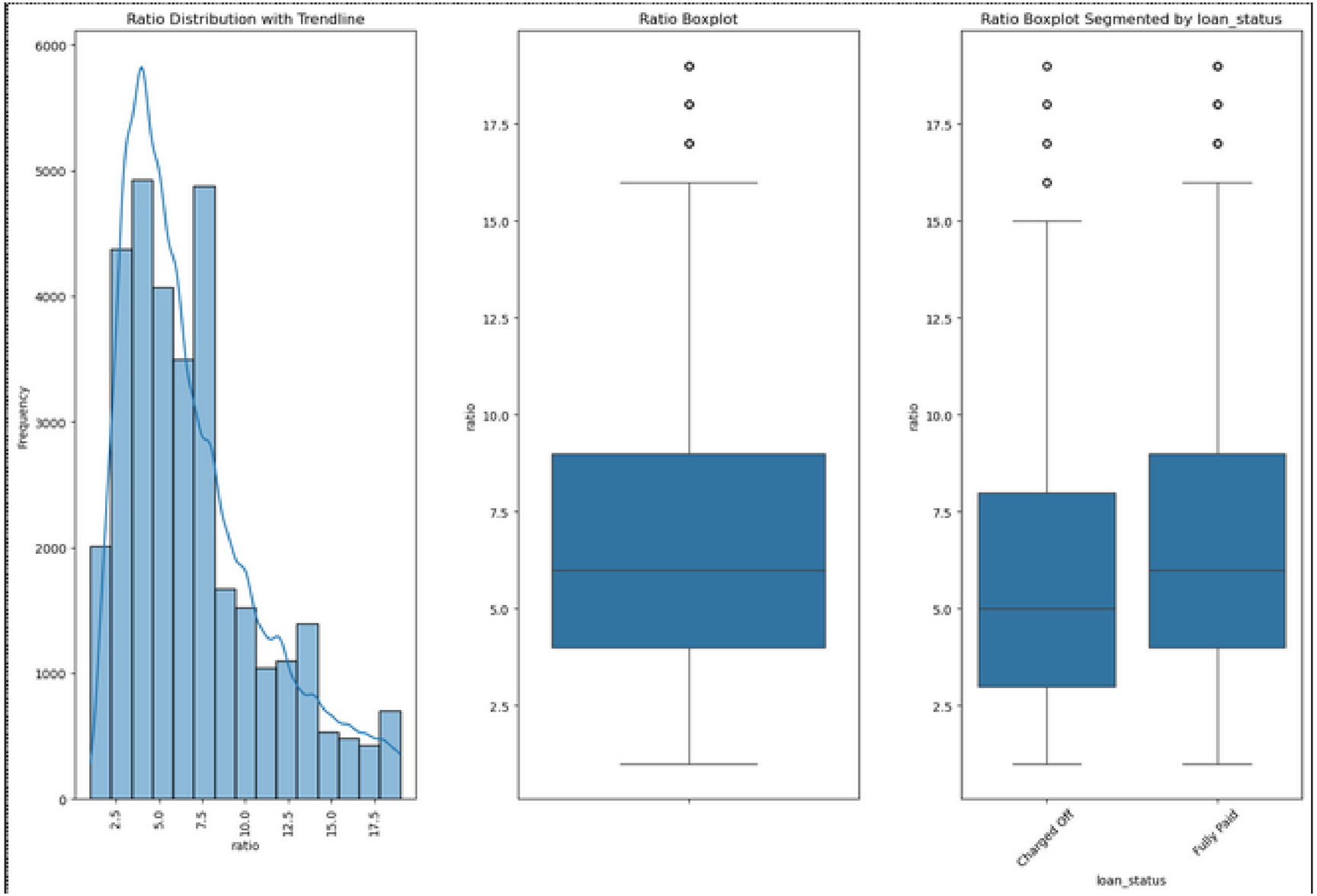
21

Ratio of annual Income to the Loan Amount Taken

Observations:

- Lower the ratio higher the chance of being default.



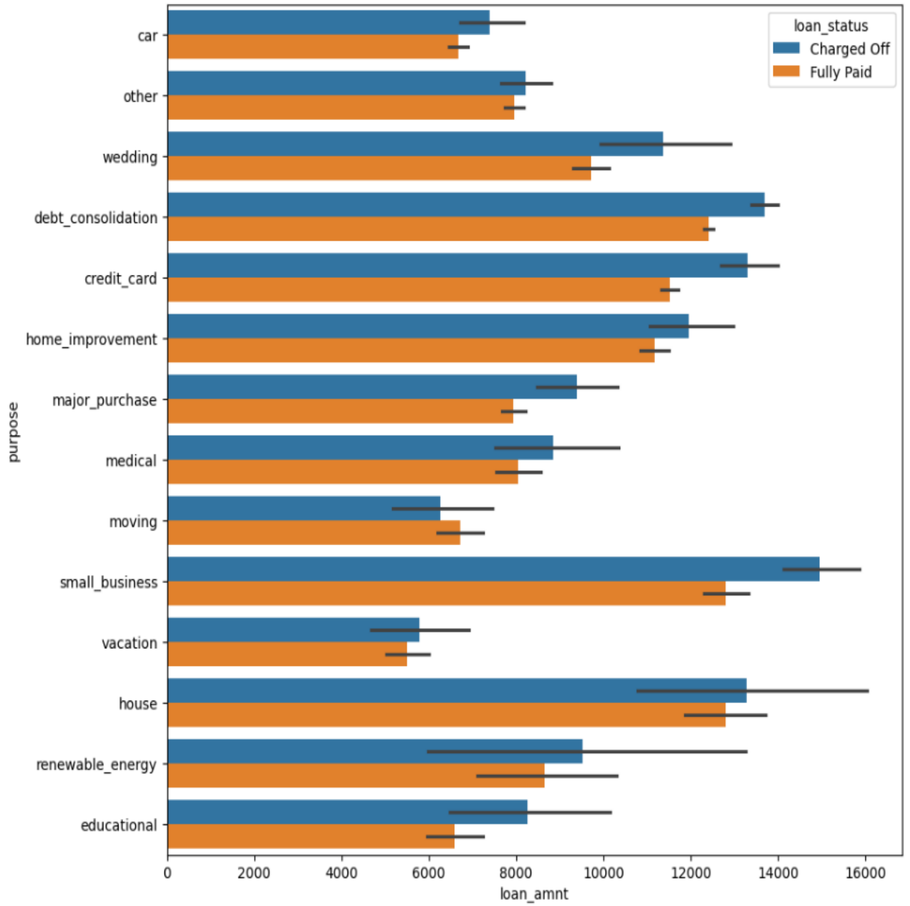

Ratio Distribution with Trendline
Frequency
ratio
Ratio Boxplot
ratio
Ratio Boxplot Segmented by loan_status
ratio
Charged Off
Fully Paid
loan_status

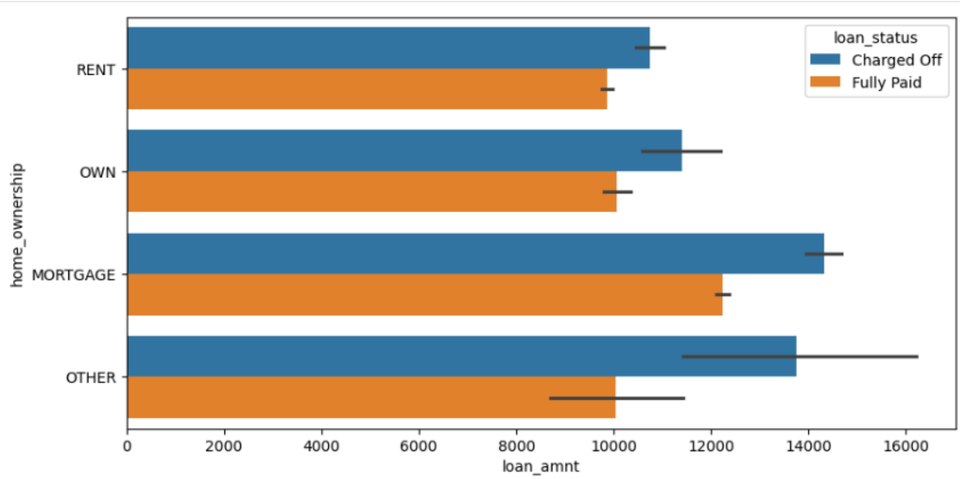
23

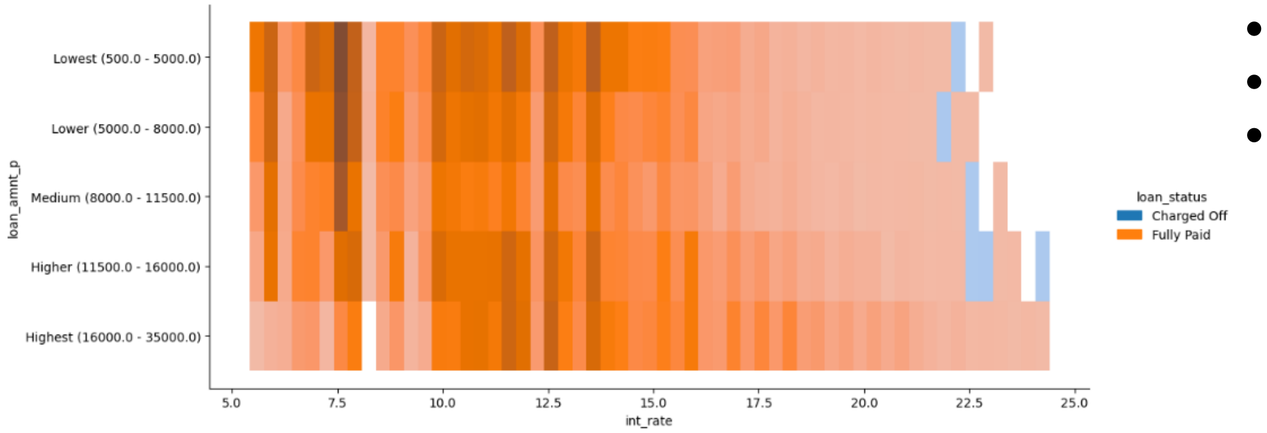
Bivariate Analysis

Analysing loan_amount with other Categorical variables

- High interest rate has higher chance of default across all loan amount groups.
- small_business Applicaiton has higher chances of default, when loan_amount is higher than 14K.
- Applicant having home ownership as MORTGAE and loan_amount greater than 14K has higher chance of defaulting.

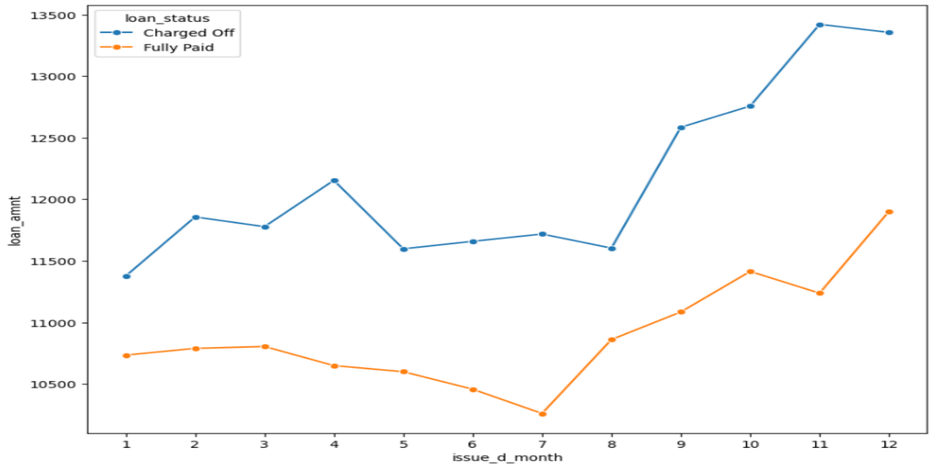


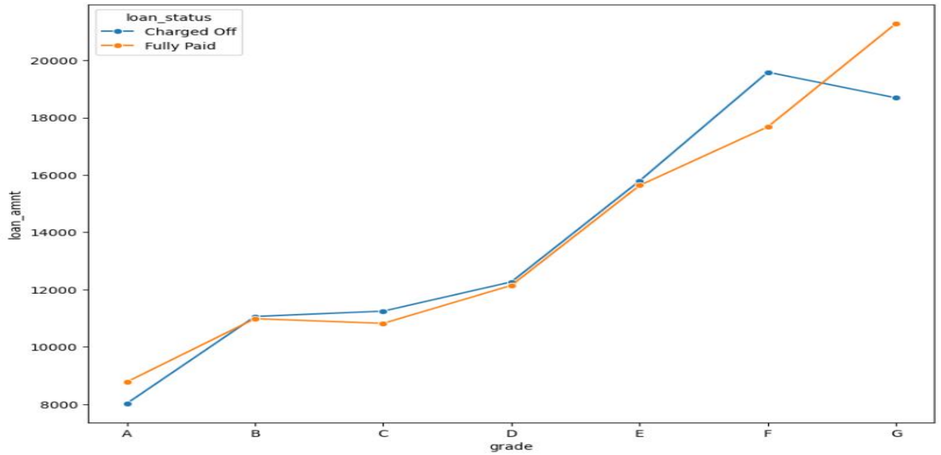


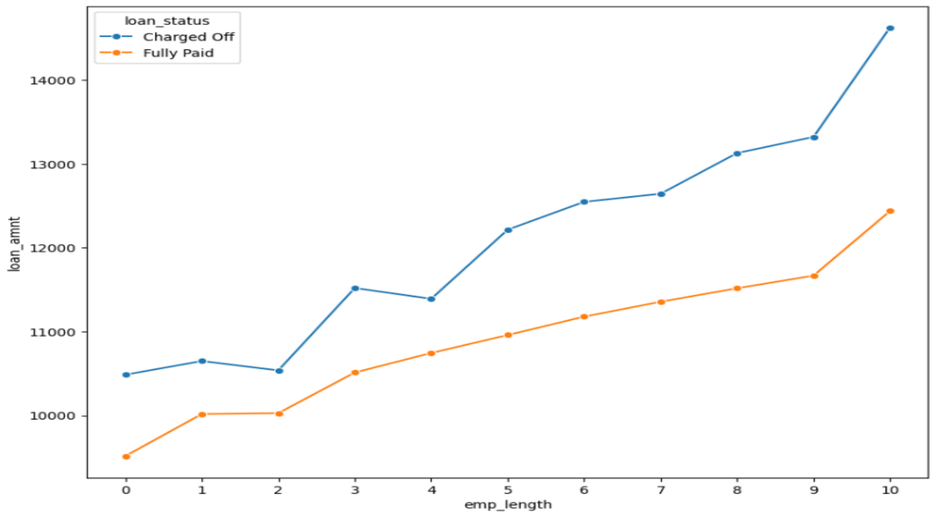


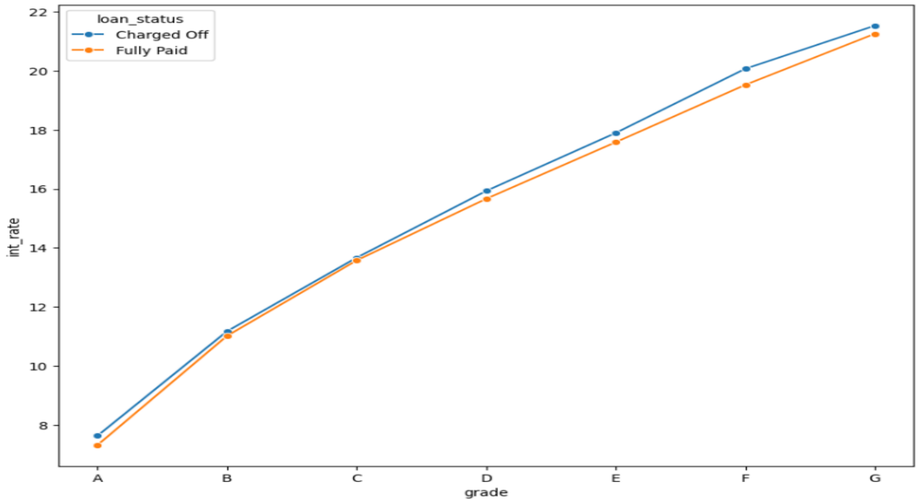
25

- Default chance is higher in December with average loan amount >13K.
- Employment length >10+ years have more loan defaults with an average loan amount 14500.
- F grade has highest defaults when average loan amount is between 15K - 20K.
- Average interest rate increases with increase in grade. G Grade has high chances of default when interest rate is above 20%.



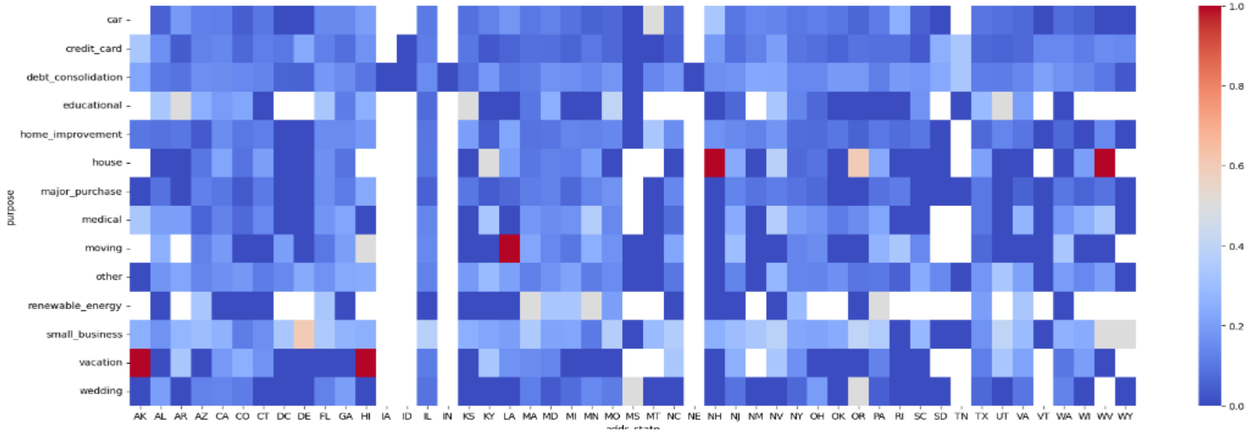


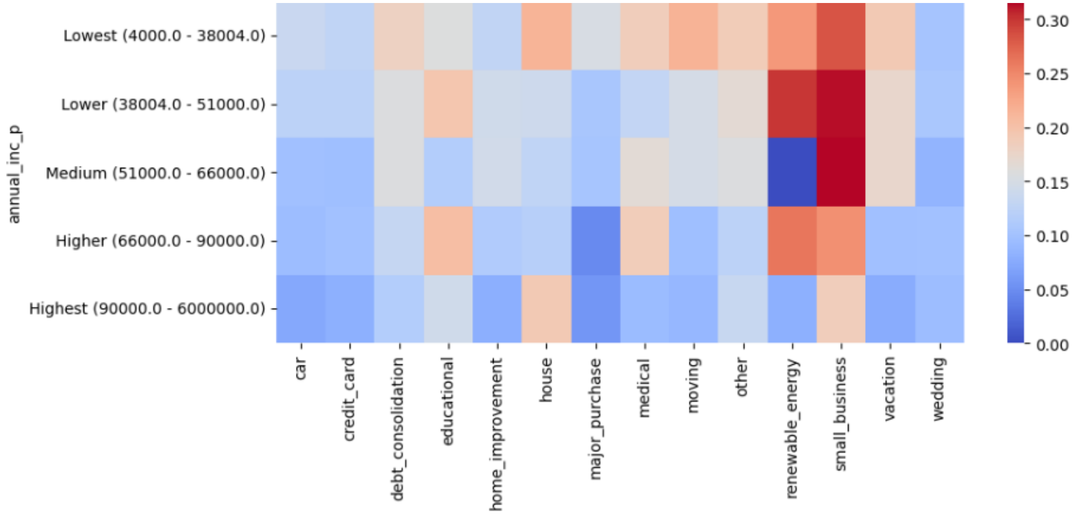


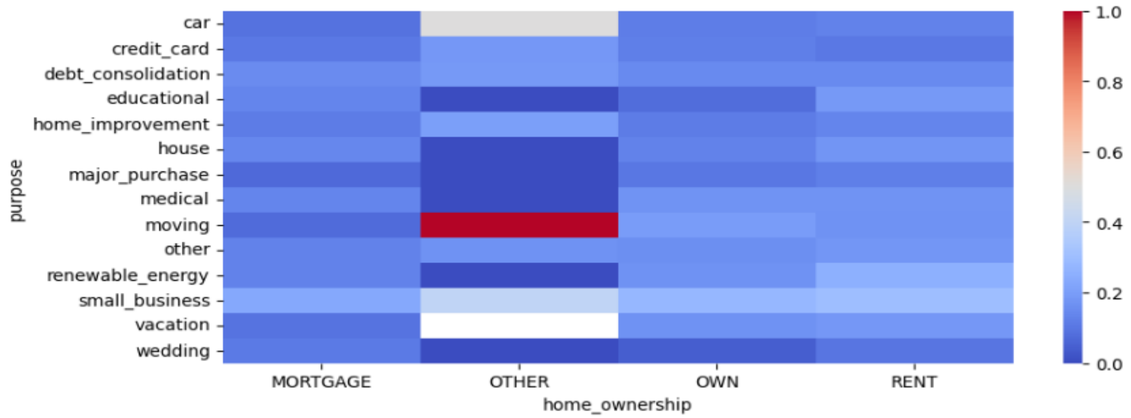


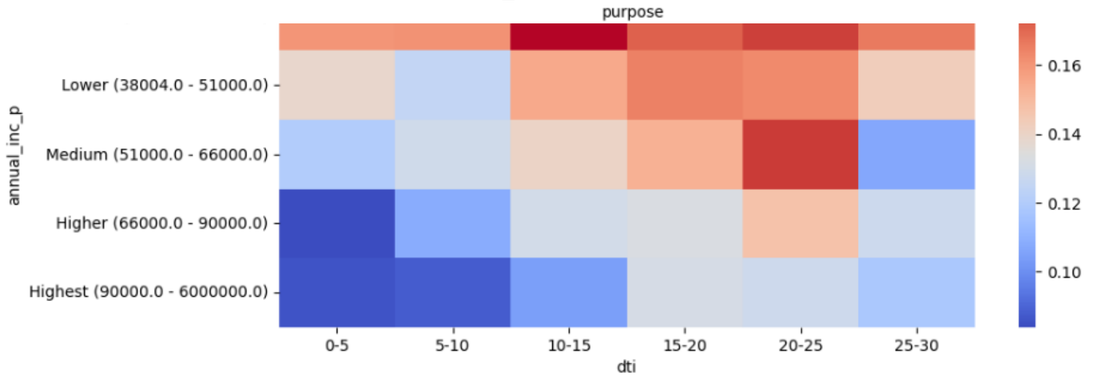
Bivariate Categorical Analysis

- Heatmap plot between address status and purpose columns shows that vacation loans in AK, HI, OR has higher defaults. House loan in NH, WV has higher defaults. Small business loans in DE, NM, WV, WY are risky.
- small business loans for lowest and medium income groups and renewable energy loans for higher income group has higher defaults.
- Medium debt-to-income group in the lowest income range has higher defaults.
- home ownership OTHER has higher default when loan purpose is moving or car.









27

Conclusions

- Income range between 0-20000 has high chances of charged off.
- Interest rate more than 16% has good chances of charged off as compared to other category interest rates.
- Those who are not owning the home is having high chances of loan defaulter.
- Those applicants having loan for small business is having high chances for loan defaults.
- High DTI value having high risk of defaults.
- Higher the Bankruptcies record higher the chance of loan defaults.
- DE States is holding highest number of loan defaults.
- The Loan applicants with loan Grade G is having highest Loan Defaults.
- Applicant having home ownership as **MORTGAE** and **Average Annual income 60K** or loan amount **more than 14K** has higher chance of defaulting.
- **small_business** loan where **avegrage loan_amount is more than 14K** has higher default.
- Verified status having loan amount **greater than 16K** has higher chance of defaulting. This also indicates issue in verification process.
- **G grade** has higher defaults when interest rate is **above 20%**. **F grade** has highest defaults when average loan amount is between **15-20K**.
- **Employment length >10+ years** has higher chances of defaults with an average loan amount 14500.
- December month has higher defaults having average loan amount 13K.
- vacation loans in AK, HI, OR is risky. House loans in NH, WV is risky. small business loans in DE, NM, WV, WY is risky.
- **small business** loans for lowest and medium income groups (4K-65K) has higher chance of defaulting.
- Medium debt-to-income group in the lowest income range has higher chances of defaulting.
- home ownership **OTHER** has high default when loan purpose is **moving or car**.
- Applicants with prb_rec/pub_rec_bankruptcies or open_acc/total_acc has higher chances of defaulting.